Chapter 3

Trends in Spatial Data Mining

Shashi Shekhar[*], *Pusheng Zhang*[*], *Yan Huang*[*], *Ranga Raju Vatsavai*[*]

[*]Department of Computer Science and Engineering, University of Minnesota
4-192, 200 Union ST SE, Minneapolis, MN 55455

Abstract:
Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial datasets. Extracting interesting and useful patterns from spatial datasets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. This chapter focuses on the unique features that distinguish spatial data mining from classical data mining. Major accomplishments and research needs in spatial data mining research are discussed.

Keywords:
Spatial Data Mining, Spatial Autocorrelation, Location Prediction, Spatial Outliers, Co-location, Spatial Statistics, Research Needs

3.1 Introduction

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining [Roddick & Spiliopoulou1999, Shekhar & Chawla2003] is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Efficient tools for extracting information from geo-spatial data are crucial to organizations which make decisions based on large spatial datasets, including NASA, the National Imagery and Mapping Agency (NIMA), the National Cancer Institute (NCI), and the United States Department of Transportation (USDOT). These organizations are spread across many application domains including ecology and environmental management, public safety, transportation, Earth science, epidemiology, and climatology.

General purpose data mining tools, such as Clementine, See5/C5.0, and Enterprise Miner, are designed to analyze large commercial databases. Although these tools were primarily designed to identify customer-buying patterns in market basket data, they have also been used in analyzing scientific and engineering data, astronomical data, multi-media data, genomic data, and web data. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation.

Specific features of geographical data that preclude the use of general purpose data mining algorithms are: i) rich data types(e.g., extended spatial objects) ii) implicit spatial relationships among the variables, iii) observations that are not independent, and iv) spatial autocorrelation among the features. In this chapter we focus on the unique features that distinguish spatial data mining from classical data mining in the following four categories: data input, statistical foundation, output patterns, and computational process. We present major accomplishments of spatial data mining research, especially regarding output patterns known as predictive models, spatial outliers, spatial co-location rules, and clusters. Finally, we identify areas of spatial data mining where further research is needed.

3.2 Data Input

The data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as points, lines, and polygons. The data inputs of spatial data mining have two distinct types of attributes: non-spatial attribute and spatial attribute. Non-spatial attributes are used to characterize non-spatial features of objects, such as name, population, and unemployment rate for a city. They are the same as the attributes used in the data inputs of classical data mining. Spatial attributes are used

to define the spatial location and extent of spatial objects [Bolstad2002]. The spatial attributes of a spatial object most often include information related to spatial locations, e.g., longitude, latitude and elevation, as well as shape.

Relationships among non-spatial objects are explicit in data inputs, e.g., arithmetic relation, ordering, is_instance_of, subclass_of, and membership_of. In contrast, relationships among spatial objects are **often implicit**, such as overlap, intersect, and behind. One possible way to deal with implicit spatial relationships is to materialize the relationships into traditional data input columns and then apply classical data mining techniques [Quinlan1993, Barnett & Lewis1994, Agrawal & Srikant1994, Jain & Dubes1988]. However, the materialization can result in loss of information. Another way to capture implicit spatial relationships is to develop models or techniques to incorporate spatial information into the spatial data mining process. We discuss a few case studies of such techniques in Section 3.4.

Non-spatial Relationship (Explicit)	Spatial Relationship (Often Implicit)
Arithmetic	Set-oriented: union, intersection, membership, $\cdots$
Ordering	Topological: meet, within, overlap, $\cdots$
Is_instance_of	Directional: North, NE, left, above, behind, $\cdots$
Subclass_of	Metric: e.g., distance, area, perimeter, $\cdots$
Part_of	Dynamic: update, create, destroy, $\cdots$
Membership_of	Shape-based and visibility

Table 3.1: Relationships among Non-spatial Data and Spatial Data

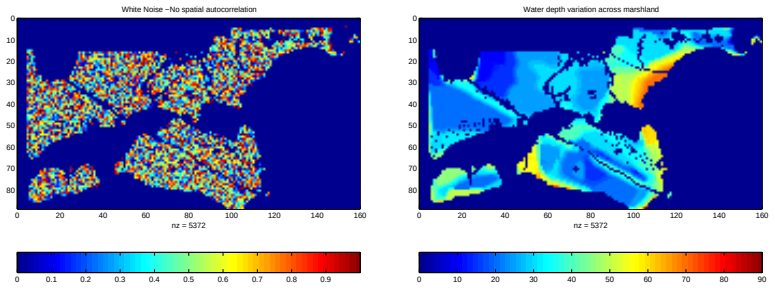
3.3 Statistical Foundation

Statistical models [Cressie1993] are often used to represent observations in terms of random variables. These models can then be used for estimation, description, and prediction based on probability theory. Spatial data can be thought of as resulting from observations on the stochastic process $Z(s)$: $s \in D$, where s is a spatial location and D is possibly a random set in a spatial framework. Here we present three spatial statistical problems one might encounter: point process, lattice, and geostatistics.

Point process: A point process is a model for the spatial distribution of the points in a point pattern. Several natural processes can be modeled as spatial point patterns, e.g., positions of trees in a forest and locations of bird habitats in a wetland. Spatial point patterns can be broadly grouped into random or non-random processes. Real point patterns are often compared with a random pattern(generated by a Poisson process) using the average distance between a point and its nearest neighbor. For a random pattern, this average distance is expected to be $\frac{1}{2*\sqrt{density}}$, where density is the average number of points per unit area. If for a real process, the computed distance falls within a certain limit, then we conclude that the pattern is generated by a random process; otherwise it is a non-random process.

Lattice: A lattice is a model for a gridded space in a spatial framework. Here the lattice refers to a countable collection of regular or irregular spatial sites related to each other via a neighborhood relationship. Several spatial statistical analyses, e.g., the spatial autoregressive model and Markov random fields, can be applied on lattice data.

Geostatistics: Geostatistics deals with the analysis of spatial continuity and weak stationarity [Cressie1993], which is an inherent characteristics of spatial data sets. Geostatistics provides a set of statistics tools, such as kriging [Cressie1993] to the interpolation of attributes at unsampled locations.



(a) Attribute with an Independent Identical Distribution

(b) Attribute with Spatial Autocorrelation

Figure 3.1: Attribute Values in Space with Independent Identical Distribution and Spatial Autocorrelation

One of the fundamental assumptions of statistical analysis is that the data samples are independently generated: like successive tosses of coin, or the rolling of a die. However, in the analysis of spatial data, the assumption about the independence of samples is generally false. In fact, spatial data tends to be highly self correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. The economies of a region tend to be similar. Changes in natural resources, wildlife, and temperature vary gradually over space. The property of like things to cluster in space is so fundamental that geographers have elevated it to the status of the first law of geography: *"Everything is related to everything else but nearby things are more related than distant things"* [Tobler1979]. In spatial statistics, an area within statistics devoted to the analysis of spatial data, this property is called **spatial autocorrelation**. For example, Figure 3.1 shows the value distributions of an attribute in a spatial framework for an independent identical distribution and a distribution with spatial autocorrelation.

Knowledge discovery techniques which ignore spatial autocorrelation typically perform poorly in the presence of spatial data. Often the spatial dependencies arise due to the inherent characteristics of the phenomena under study, but in particular they arise due to the fact that the spatial resolution of imag-

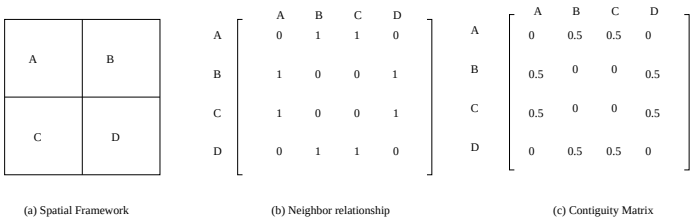


Figure 3.2: A Spatial Framework and Its Four-neighborhood Contiguity Matrix.

ing sensors are finer than the size of the object being observed. For example, remote sensing satellites have resolutions ranging from 30 meters (e.g., the Enhanced Thematic Mapper of the Landsat 7 satellite of NASA) to one meter (e.g., the IKONOS satellite from SpaceImaging), while the objects under study (e.g., Urban, Forest, Water) are often much larger than 30 meters. As a result, per-pixel-based classifiers, which do not take spatial context into account, often produce classified images with *salt and pepper* noise. These classifiers also suffer in terms of classification accuracy.

The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix. A simple contiguity matrix may represent a neighborhood relationship defined using adjacency, Euclidean distance, etc. Example definitions of neighborhood using adjacency include a four-neighborhood and an eight-neighborhood. Given a gridded spatial framework, a four-neighborhood assumes that a pair of locations influence each other if they share an edge. An eight-neighborhood assumes that a pair of locations influence each other if they share either an edge or a vertex.

Figure 3.2(a) shows a gridded spatial framework with four locations, A, B, C, and D. A binary matrix representation of a four-neighborhood relationship is shown in Figure 3.2(b). The row-normalized representation of this matrix is called a contiguity matrix, as shown in Figure 3.2(c). Other contiguity matrices can be designed to model neighborhood relationships based on distance. The essential idea is to specify the pairs of locations that influence each other along with the relative intensity of interaction. More general models of spatial relationships using cliques and hypergraphs are available in the literature [Warrender & Augusteijn1999]. In spatial statistics, spatial autocorrelation is quantified using measures such as Ripley's K-function and Moran's I [Cressie1993].

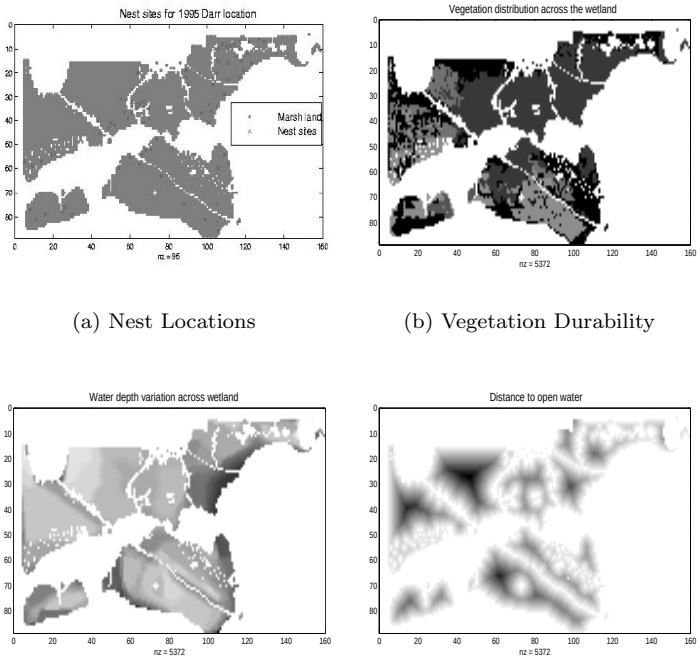
3.4 Output Patterns

In this section, we present case studies of four important output patterns for spatial data mining: predictive models, spatial outliers, spatial co-location rules, and spatial clustering.

3.4.1 Predictive Models

The prediction of events occurring at particular geographic locations is very important in several application domains. Examples of problems which require location prediction include crime analysis, cellular networking, and natural disasters such as fires, floods, droughts, vegetation diseases, and earthquakes. In this section we provide two spatial data mining techniques for predicting locations, namely the Spatial Autoregressive Model (SAR) and Markov Random Fields (MRF).

An Application Domain We begin by introducing an example to illustrate the different concepts related to location prediction in spatial data mining. We are given data about two wetlands, named Darr and Stubble, on the shores of Lake Erie in Ohio USA in order to *predict* the spatial distribution of a marsh-breeding bird, the red-winged blackbird (*Agelaius phoeniceus*). The data was collected from April to June in two successive years, 1995 and 1996.



(a) Nest Locations

(b) Vegetation Durability

(c) Water Depth

(d) Distance to Open Water

Figure 3.3: (a) Learning dataset: The geometry of the Darr wetland and the locations of the nests, (b) The spatial distribution of *vegetation durability* over the marshland, (c) The spatial distribution of *water depth*, and (d) The spatial distribution of *distance to open water.*

A uniform grid was imposed on the two wetlands and different types of measurements were recorded at each cell or pixel. In total, the values of seven attributes were recorded at each cell. Domain knowledge is crucial in deciding which attributes are important and which are not. For example, *Vegetation Durability* was chosen over *Vegetation Species* because specialized knowledge about the bird-nesting habits of the red-winged blackbird suggested that the choice of nest location is more dependent on plant structure, plant resistance to wind, and wave action than on the plant species.

An important goal is to build a model for predicting the location of bird nests in the wetlands. Typically, the model is built using a portion of the data, called the learning or training data, and then tested on the remainder of the data, called the testing data. In this study we build a model using the 1995 Darr wetland data and then tested it 1995 Stubble wetland data. In the learning data, all the attributes are used to build the model and in the training data, one value is hidden, in our case the location of the nests. Using knowledge gained from the 1995 Darr data and the value of the independent attributes in the test data, we want to predict the location of the nests in 1995 Stubble data.

Modeling Spatial Dependencies Using the SAR and MRF Models Several previous studies [Jhung & Swain1996], [Solberg, Taxt, & Jain1996] have shown that the modeling of spatial dependency (often called context) during the classification process improves overall classification accuracy. Spatial context can be defined by the relationships between spatially adjacent pixels in a small neighborhood. In this section, we present two models to model spatial dependency: the spatial autoregressive model(SAR) and Markov random field(MRF)-based Bayesian classifiers.

Spatial Autoregressive Model The spatial autoregressive model decomposes a classifier $\hat{f}_C$ into two parts, namely spatial autoregression and logistic transformation. We first show how spatial dependencies are modeled using the framework of logistic regression analysis. In the spatial autoregression model, the spatial dependencies of the error term, or, the dependent variable, are directly modeled in the regression equation[Anselin1988]. If the dependent values y_i are related to each other, then the regression equation can be modified as

$$y = \rho W y + X\beta + \epsilon. \tag{3.1}$$

Here W is the neighborhood relationship contiguity matrix and ρ is a parameter that reflects the strength of the spatial dependencies between the elements of the dependent variable. After the correction term $\rho W y$ is introduced, the components of the residual error vector ϵ are then assumed to be generated from independent and identical standard normal distributions. As in the case of classical regression, the SAR equation has to be transformed via the logistic function for binary dependent variables.

We refer to this equation as the Spatial Autoregressive Model (SAR). Notice that when $\rho = 0$, this equation collapses to the classical regression model. The benefits of modeling spatial autocorrelation are many: The residual error will have much lower spatial autocorrelation (i.e., systematic variation). With the

proper choice of W, the residual error should, at least theoretically, have no systematic variation. If the spatial autocorrelation coefficient is statistically significant, then SAR will quantify the presence of spatial autocorrelation. It will indicate the extent to which variations in the dependent variable (y) are explained by the average of neighboring observation values. Finally, the model will have a better fit, (i.e., a higher R-squared statistic).

Markov Random Field-based Bayesian Classifiers Markov random field-based Bayesian classifiers estimate the classification model $\hat{f}_C$ using MRF and Bayes' rule. A set of random variables whose interdependency relationship is represented by an undirected graph (i.e., a symmetric neighborhood matrix) is called a Markov Random Field [Li1995]. The Markov property specifies that a variable depends only on its neighbors and is independent of all other variables. The location prediction problem can be modeled in this framework by assuming that the class label, $l_i = f_C(s_i)$, of different locations, s_i, constitutes an MRF. In other words, random variable l_i is independent of l_j if $W(s_i, s_j) = 0$.

The Bayesian rule can be used to predict l_i from feature value vector X and neighborhood class label vector L_i as follows:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X)} \tag{3.2}$$

The solution procedure can estimate $Pr(l_i|L_i)$ from the training data, where L_i denotes a set of labels in the neighborhood of s_i excluding the label at s_i, by examining the ratios of the frequencies of class labels to the total number of locations in the spatial framework. $Pr(X|l_i, L_i)$ can be estimated using kernel functions from the observed values in the training dataset. For reliable estimates, even larger training datasets are needed relative to those needed for the Bayesian classifiers without spatial context, since we are estimating a more complex distribution. An assumption on $Pr(X|l_i, L_i)$ may be useful if the training dataset available is not large enough. A common assumption is the uniformity of influence from all neighbors of a location. For computational efficiency it can be assumed that only local explanatory data $X(s_i)$ and neighborhood label L_i are relevant in predicting class label $l_i = f_C(s_i)$. It is common to assume that all interaction between neighbors is captured via the interaction in the class label variable. Many domains also use specific parametric probability distribution forms, leading to simpler solution procedures. In addition, it is frequently easier to work with a Gibbs distribution specialized by the locally defined MRF through the Hammersley-Clifford theorem [Besag1974].

A more detailed theoretical and experimental comparison of these methods can be found in [Shekhar *et al.*2002]. Although MRF and SAR classification have different formulations, they share a common goal, estimating the posterior probability distribution: $p(l_i|X)$. However, the posterior for the two models is computed differently with different assumptions. For MRF the posterior is computed using Bayes' rule. On the other hand, in logistic regression, the posterior distribution is directly fit to the data. One important difference between logistic regression and MRF is that logistic regression assumes no de-

pendence on neighboring classes. Logistic regression and logistic SAR models belong to a more general exponential family. The exponential family is given by $Pr(u|v) = e^{A(\theta_v)+B(u,\pi)+\theta_v^T u}$ where u, v are location and label respectively. This exponential family includes many of the common distributions such as Gaussian, Binomial, Bernoulli, and Poisson as special cases.

Experiments were carried out on the Darr and Stubble wetlands to compare classical regression, SAR, and the MRF-based Bayesian classifiers. The results showed that the MRF models yield better spatial and classification accuracies over SAR in the prediction of the locations of bird nests. We also observed that SAR predictions are extremely localized, missing actual nests over a large part of the marsh lands.

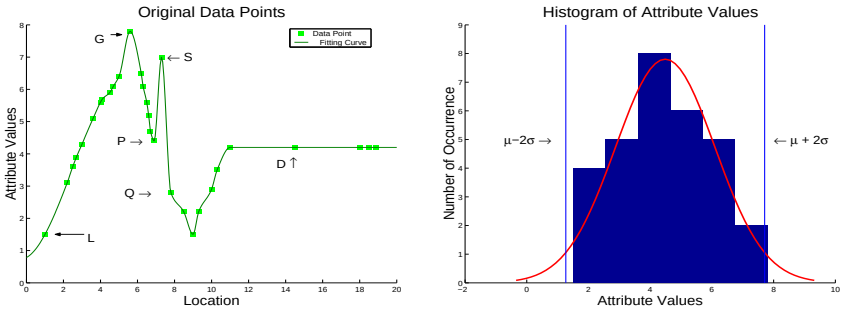
3.4.2 Spatial Outliers

Outliers have been informally defined as observations in a dataset which appear to be inconsistent with the remainder of that set of data [Barnett & Lewis1994], or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [Hawkins1980]. The identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as credit card fraud, athlete performance analysis, voting irregularity, and severe weather prediction. This section focuses on spatial outliers, i.e., observations which appear to be inconsistent with their neighborhoods. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases, including transportation, ecology, public safety, public health, climatology, and location-based services.

A spatial outlier is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. Informally, a spatial outlier is a local instability (in values of non-spatial attributes) or a spatially referenced object whose non-spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population. For example, a new house in an old neighborhood of a growing metropolitan area is a spatial outlier based on the non-spatial attribute house age.

Illustrative Examples We use an example to illustrate the differences among global and spatial outlier detection methods. In Figure 3.4(a), the X-axis is the location of data points in one-dimensional space; the Y-axis is the attribute value for each data point. Global outlier detection methods ignore the spatial location of each data point and fit the distribution model to the values of the non-spatial attribute. The outlier detected using this approach is the data point G, which has an extremely high attribute value 7.9, exceeding the threshold of $\mu + 2\sigma = 4.49 + 2 * 1.61 = 7.71$, as shown in Figure 3.4(b). This test assumes a normal distribution for attribute values. On the other hand, S is a spatial outlier whose observed value is significantly different than its neighbors P and Q.

Tests for Detecting Spatial Outliers Tests to detect spatial outliers



(a) An Example Dataset

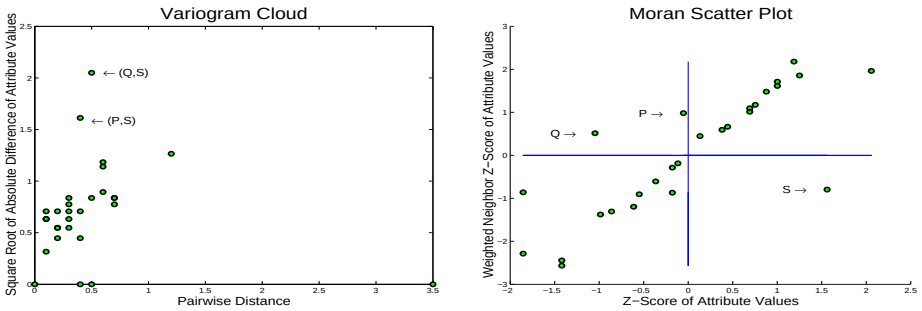
(b) Histogram

Figure 3.4: A Dataset for Outlier Detection.

separate spatial attributes from non-spatial attributes. Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Spatial statistics literature provides two kinds of bi-partite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests, which are based on the visualization of spatial data, highlight spatial outliers. Example methods include variogram clouds and Moran scatterplots. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data. Scatterplots [Anselin1994] are a representative technique from the quantitative family.

A variogram-cloud [Cressie1993] displays data points related by neighborhood relationships. For each pair of locations, the square-root of the absolute difference between attribute values at the locations versus the Euclidean distance between the locations are plotted. In datasets exhibiting strong spatial dependence, the variance in the attribute differences will increase with increasing distance between locations. Locations that are near to one another, but with large attribute differences, might indicate a spatial outlier, even though the values at both locations may appear to be reasonable when examining the dataset non-spatially. Figure 3.5(a) shows a variogram cloud for the example dataset shown in Figure 3.4(a). This plot shows that two pairs (P, S) and (Q, S) on the left hand side lie above the main group of pairs, and are possibly related to spatial outliers. The point S may be identified as a spatial outlier since it occurs in both pairs (Q, S) and (P, S). However, graphical tests of spatial outlier detection are limited by the lack of precise criteria to distinguish spatial outliers. In addition, a variogram cloud requires non-trivial post-processing of highlighted pairs to separate spatial outliers from their neighbors, particularly when multiple outliers are present, or density varies greatly.

A Moran scatterplot [Anselin1995] is a plot of normalized attribute value $(Z[f(i)] = \frac{f(i)-\mu_f}{\sigma_f})$ against the neighborhood average of normalized attribute



(a) Variogram cloud

(b) Moran scatterplot

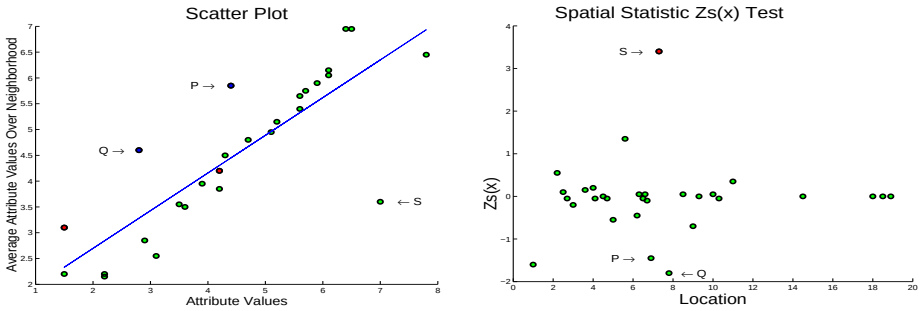
Figure 3.5: Variogram Cloud and Moran Scatterplot to Detect Spatial Outliers.

values ($W \cdot Z$), where W is the row-normalized (i.e., $\sum_j W_{ij} = 1$) neighborhood matrix, (i.e., $W_{ij} > 0$ iff neighbor(i, j)). The upper left and lower right quadrants of Figure 3.5(b) indicate a spatial association of dissimilar values: low values surrounded by high value neighbors(e.g., points P and Q), and high values surrounded by low values (e.g,. point S). Thus we can identify points(nodes) that are surrounded by unusually high or low value neighbors. These points can be treated as spatial outliers.

A scatterplot [Anselin1994] shows attribute values on the X-axis and the average of the attribute values in the neighborhood on the Y-axis. A least square regression line is used to identify spatial outliers. A scatter sloping upward to the right indicates a positive spatial autocorrelation (adjacent values tend to be similar); a scatter sloping upward to the left indicates a negative spatial autocorrelation. The residual is defined as the vertical distance (Y-axis) between a point P with location (X_p, Y_p) to the regression line $Y = mX+b$, that is, residual $\epsilon = Y_p - (mX_p + b)$. Cases with standardized residuals, $\epsilon_{standard} = \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon}$, greater than 3.0 or less than -3.0 are flagged as possible spatial outliers, where μ_ϵ and σ_ϵ are the mean and standard deviation of the distribution of the error term ϵ. In Figure 3.6(a), a scatterplot shows the attribute values plotted against the average of the attribute values in neighboring areas for the dataset in Figure 3.4(a). The point S turns out to be the farthest from the regression line and may be identified as a spatial outlier.

A location (sensor) is compared to its neighborhood using the function $S(x) = [f(x) - E_{y \in N(x)}(f(y))]$, where $f(x)$ is the attribute value for a location x, $N(x)$ is the set of neighbors of x, and $E_{y \in N(x)}(f(y))$ is the average attribute value for the neighbors of x [Shekhar, Lu, & Zhang2003]. The statistic function $S(x)$ denotes the difference of the attribute value of a sensor located at x and the average attribute value of $x's$ neighbors.

Spatial statistic $S(x)$ is normally distributed if the attribute value $f(x)$ is



(a) Scatterplot

(b) Spatial statistic $Z_{s(x)}$

Figure 3.6: Scatterplot and Spatial Statistic $Z_{s(x)}$ to Detect Spatial Outliers.

normally distributed. A popular test for detecting spatial outliers for normally distributed $f(x)$ can be described as follows: Spatial statistic $Z_{s(x)} = |\frac{S(x)-\mu_s}{\sigma_s}| > \theta$. For each location x with an attribute value $f(x)$, the $S(x)$ is the difference between the attribute value at location x and the average attribute value of $x's$ neighbors, μ_s is the mean value of $S(x)$, and σ_s is the value of the standard deviation of $S(x)$ over all stations. The choice of θ depends on a specified confidence level. For example, a confidence level of 95 percent will lead to $\theta \approx 2$.

Figure 3.6(b) shows the visualization of the spatial statistic method described above. The X-axis is the location of data points in one-dimensional space; the Y-axis is the value of spatial statistic $Z_{s(x)}$ for each data point. We can easily observe that point S has a $Z_{s(x)}$ value exceeding 3, and will be detected as a spatial outlier. Note that the two neighboring points P and Q of S have $Z_{s(x)}$ values close to -2 due to the presence of spatial outliers in their neighborhoods.

3.4.3 Spatial Co-location Rules

Boolean spatial features are geographic object types which are either present or absent at different locations in a two dimensional or three dimensional metric space, e.g., the surface of the Earth. Examples of boolean spatial features include plant species, animal species, road types, cancers, crime, and business types. Co-location patterns represent the subsets of the boolean spatial features whose instances are often located in close geographic proximity. Examples include symbiotic species, e.g., Nile crocodile and Egyptian plover in ecology, and frontage roads and highways in metropolitan road maps.

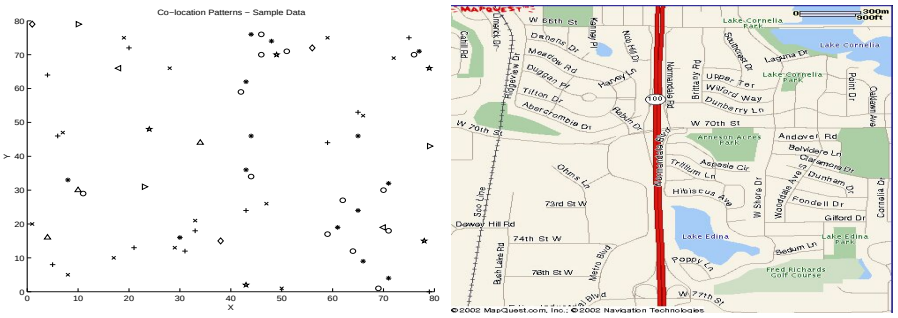


Figure 3.7: (a) Illustration of Point Spatial Co-location Patterns. Shapes represent different spatial feature types. Spatial features in sets {'+', '×'} and {'o', '*'} tend to be located together. (b) Illustration of Line String Co-location Patterns. Highways, e.g., Hwy100, and frontage roads, e.g., Normandale Road, are co-located.

Co-location rules are models to infer the presence of boolean spatial features in the neighborhood of instances of other boolean spatial features. For example, "Nile Crocodiles → Egyptian Plover" predicts the presence of Egyptian Plover birds in areas with Nile Crocodiles. Figure 3.7(a) shows a dataset consisting of instances of several boolean spatial features, each represented by a distinct shape. A careful review reveals two co-location patterns, i.e., ('+','×') and ('o','*').

Co-location rule discovery is a process to identify co-location patterns from large spatial datasets with a large number of boolean features. The spatial co-location rule discovery problem looks similar to, but, in fact, is very different from the association rule mining problem [Agrawal & Srikant1994] because of the lack of transactions. In market basket datasets, transactions represent sets of item types bought together by customers. The support of an association is defined to be the fraction of transactions containing the association. Association rules are derived from all the associations with support values larger than a user given threshold. The purpose of mining association rules is to identify frequent item sets for planning store layouts or marketing campaigns. In the spatial co-location rule mining problem, transactions are often not explicit. The transactions in market basket analysis are independent of each other. Transactions are disjoint in the sense of not sharing instances of item types. In contrast, the instances of Boolean spatial features are embedded in a continuous space and share a variety of spatial relationships (e.g., neighbor) with each other.

Co-location Rule Approaches Approaches to discovering co-location rules in the literature can be categorized into three classes, namely spatial statistics, association rules, and the event centric approach. Spatial statistics-based approaches use measures of spatial correlation to characterize the relationship between different types of spatial features using the cross K function with Monte

Carlo simulation and quadrat count analysis [Cressie1993]. Computing spatial correlation measures for all possible co-location patterns can be computationally expensive due to the exponential number of candidate subsets given a large collection of spatial boolean features.

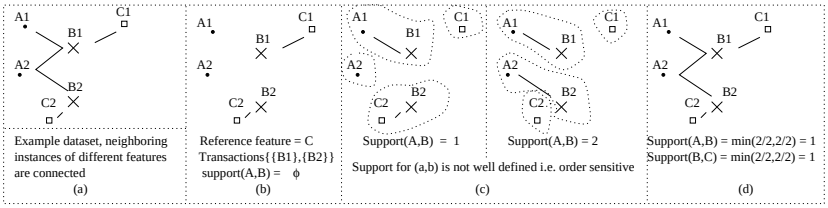


Figure 3.8: Example to Illustrate Different Approaches to Discovering Co-location Patterns (a) Example dataset. (b) Data partition approach. Support measure is ill-defined and order sensitive (c) Reference feature centric model (d) Event centric model

Association rule-based approaches focus on the creation of transactions over space so that an *apriori* like algorithm [Agrawal & Srikant1994] can be used. Transactions over space can use a reference-feature centric [Koperski & Han1995] approach or a data-partition [Morimoto2001] approach. The reference feature centric model is based on the choice of a reference spatial feature and is relevant to application domains focusing on a specific boolean spatial feature, e.g., incidence of cancer. Domain scientists are interested in finding the co-locations of other task relevant features (e.g., asbestos) to the reference feature. A specific example is provided by the spatial association rule presented in [Koperski & Han1995]. Transactions are created around instances of one user-specified reference spatial feature. The association rules are derived using the *apriori* [Agrawal & Srikant1994] algorithm. The rules found are all related to the reference feature. For example, consider the spatial dataset in Figure 3.8(a) with three feature types, A, B and C. Each feature type has two instances. The neighbor relationships between instances are shown as edges. Co-locations (A, B) and (B, C) may be considered to be frequent in this example. Figure 3.8(b) shows transactions created by choosing C as the reference feature. Co-location (A, B) will not be found since it does not involve the reference feature.

Defining transactions by a data-partition approach [Morimoto2001] defines transactions by dividing spatial datasets into disjoint partitions. There may be many distinct ways of partitioning the data, each yielding a distinct set of transactions, which in turn yields different values of support of a given co-location. Figure 3.8 (c) shows two possible partitions for the dataset of Figure 3.8 (a), along with the supports for co-location (A, B).

The event centric model finds subsets of spatial features likely to occur in a neighborhood around instances of given subsets of event types. For example, let us determine the probability of finding at least one instance of feature type B in the neighborhood of an instance of feature type A in Figure 3.8 (a). There are two instances of type A and both have some instance(s) of type B

Table 3.2: Interest Measures for Different Models

Model	Items	Transactions defined by	Interest measures for $C_1 \rightarrow C_2$	
			Prevalence	Conditional probability
reference feature centric	predicates on reference and relevant features	instances of reference feature C_1 and C_2 involved with	fraction of instance of reference feature with $C_1 \cup C_2$	$Pr(C_2$ is true for an instance of reference features given C_1 is true for that instance of reference feature)
data partitioning	boolean feature types	a partitioning of spatial dataset	fraction of partitions with $C_1 \cup C_2$	$Pr(C_2$ in a partition given C_1 in that partition)
event centric	boolean feature types	neighborhoods of instances of feature types	participation index of $C_1 \cup C_2$	$Pr(C_2$ in a neighborhood of $C_1)$

in their neighborhoods. The conditional probability for the co-location rule is: *spatial feature A at location* $l \rightarrow$ *spatial feature type B in neighborhood is 100%.* This yields a well-defined prevalence measure(i.e., support) without the need for transactions. Figure 3.8 (d) illustrates that our approach will identify both (A, B) and (B, C) as frequent patterns.

Prevalence measures and conditional probability measures, called interest measures, are defined differently in different models, as summarized in Table 3.2. The reference feature centric and data partitioning models "materialize" transactions and thus can use traditional support and confidence measures. The event centric approach defined new transaction free measures, e.g., the participation index (see [Shekhar & Huang2001] for details).

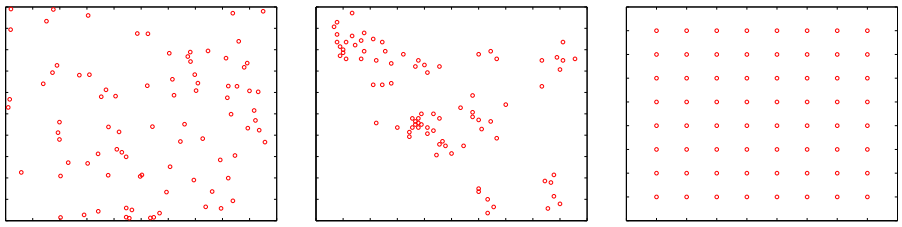
3.4.4 Spatial Clustering

Spatial clustering is a process of grouping a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. For example, clustering is used to determine the "hot spots" in crime analysis and disease tracking. Hot spot analysis is the process of finding unusually dense event clusters across time and space. Many criminal justice agencies are exploring the benefits provided by computer technologies to identify crime hot spots in order to take preventive strategies such as deploying saturation patrols in hot spot areas.

Spatial clustering can be applied to group similar spatial objects together; the implicit assumption is that patterns in space tend to be grouped rather than randomly located. However, the statistical significance of spatial clusters should be measured by testing the assumption in the data. The test is critical before proceeding with any serious clustering analyses.

Complete Spatial Randomness, Cluster, and Decluster In spatial statistics, the standard against which spatial point patterns are often compared is a completely spatially random point process, and departures indicate that the pattern is not distributed randomly in space. *Complete spatial randomness (CSR)* [Cressie1993] is synonymous with a homogeneous Poisson process. The patterns of the process are independently and uniformly distributed over space, i.e., the patterns are equally likely to occur anywhere and do not interact with each other. However, patterns generated by a non-random process can be either cluster patterns(aggregated patterns) or decluster patterns(uniformly spaced patterns).

To illustrate, Figure 3.9 shows realizations from a completely spatially random process, a spatial cluster process, and a spatial decluster process (each conditioned to have 80 points) in a square. Notice in Figure 3.9 (a) that the complete spatial randomness pattern seems to exhibit some clustering. This is not an unrepresentive realization, but illustrates a well-known property of homogeneous Poisson processes: event-to-nearest-event distances are proportional to χ^2_2 random variables, whose densities have a substantial amount of probability near zero [Cressie1993]. Spatial clustering is more statistically significant when the data exhibit a cluster pattern rather than a CSR pattern or decluster pattern.



(a) CSR Pattern (b) Cluster Pattern (c) Decluster Pattern

Figure 3.9: Illustration of CSR, Cluster, and Decluster Patterns

Several statistical methods can be applied to quantify deviations of patterns from a complete spatial randomness point pattern [Cressie1993]. One type of descriptive statistics is based on quadrats (i.e., well defined area, often rectangle in shape). Usually quadrats of random location and orientations in the quadrats are counted, and statistics derived from the counters are computed. Another type of statistics is based on distances between patterns; one such type is Ripley's K-function [Cressie1993].

After the verification of the statistical significance of the spatial clustering, classical clustering algorithms [Han, Kamber, & Tung2001] can be used to discover interesting clusters.

3.5 Computational Process

Many generic algorithmic strategies have been generalized to apply to spatial data mining. For example, as shown in Table 3.3, algorithmic strategies, such as divide-and-conquer, filter-and-refine, ordering, hierarchical structure, and parameter estimation, have been used in spatial data mining.

Generic	Spatial Data Mining
Divide-and-Conquer	Space Partitioning
Filter-and-Refine	Minimum-Bounding-Rectangle(MBR)
Ordering	Plane Sweeping, Space Filling Curves
Hierarchical Structures	Spatial Index, Tree Matching
Parameter Estimation	Parameter estimation with spatial autocorrelation

Table 3.3: Algorithmic Strategies in Spatial Data Mining

In spatial data mining, spatial autocorrelation and low dimensionality in space(e.g., 2-3) provide more opportunities to improve computational efficiency than classical data mining. NASA Earth observation systems currently generate a large sequence of global snapshots of the Earth, including various atmospheric, land, and ocean measurements such as sea surface temperature, pressure, precipitation, and net primary production. Each climate attribute in a location has a sequence of observations at different time slots, e.g., a collection of monthly temperatures from 1951-2000 in Minneapolis. Finding locations where climate attributes are highly correlated is frequently used to retrieve interesting relationships among spatial objects of Earth science data. For example, such queries are used to identify the land locations whose climate is severely affected by El Nino. However, such correlation-based queries are computationally expensive due to the large number of spatial points, e.g., more than 250k spatial cells on the Earth at a 0.5 degree by 0.5 degree resolution, and the high dimensionality of sequences, e.g., 600 for the 1951-2000 monthly temperature data.

A spatial indexing approach proposed by [Zhang *et al.*2003] exploits spatial autocorrelation to facilitate correlation-based queries. The approach groups similar time series together based on spatial proximity and constructs a search tree. The queries are processed using the search tree in a filter-and-refine style at the group level instead of at the time series level. Algebraic analyses using cost models and experimental evaluations showed that the proposed approach saves a large portion of computational cost, ranging from 40% to 98%(see [Zhang *et al.*2003] for details).

3.6 Research Needs

In this section, we discuss some areas where further research is needed in spatial data mining.

• *Comparison of classical data mining techniques with spatial data mining techniques*
As we discussed in Section 3.2, relationships among spatial objects are often implicit. It is possible to materialize the implicit relationships into traditional data input columns and then apply classical data mining techniques[Quinlan1993, Barnett & Lewis1994, Agrawal & Srikant1994, Jain & Dubes1988]. Another way to deal with implicit relationships is to use specialized spatial data mining techniques, e.g., the spatial autoregression and co-location mining. However, existing literature does not provide guidance regarding the choice between classical data mining techniques and spatial data mining techniques to mine spatial data. Therefore new research is needed to compare the two sets of approaches in effectiveness and computational efficiency.

• *Modeling semantically rich spatial properties, such as topology*
The spatial relationship among locations in a spatial framework is often modeled via a contiguity matrix using a neighborhood relationship defined using adjacency and distance. However, spatial connectivity and other complex spatial topological relationships in spatial networks are difficult to model using the continuity matrix. Research is needed to evaluate the value of enriching the continuity matrix beyond the neighborhood relationship.

• *Statistical interpretation models for spatial patterns*
Spatial patterns, such as spatial outliers and co-location rules, are identified in the spatial data mining process using unsupervised learning methods. There is a need for an independent measure of the statistical significance of such spatial patterns. For example, we may compare the co-location model with dedicated spatial statistical measures, such as Ripley's K-function, characterize the distribution of the participation index interest measure under spatial complete randomness using Monte Carlo simulation, and develop a statistical interpretation of co-location rules to compare the rules with other patterns in unsupervised learning.

Another challenge is the estimation of the detailed spatial parameters in a statistical model. Research is needed to design effective estimation procedures for the continuity matrices used in the spatial autoregressive model and Markov random field-based Bayesian classifiers from learning samples.

• *Spatial interest measures*
The interest measures of patterns in spatial data mining are different from those in classical data mining, especially regarding the four important output patterns shown in Table 3.4.

For a two-class problem, the standard way to measure classification accuracy is to calculate the percentage of correctly classified objects. However, this measure may not be the most suitable in a spatial context. *Spatial accuracy*−how far the predictions are from the actuals−is equally important in this application domain due to the effects of the discretizations of a continuous wetland into discrete pixels, as shown in Figure 3.10. Figure 3.10(a) shows the actual loca-

	Classical Data Mining	Spatial Data Mining
Predictive Model	Classification accuracy	Spatial accuracy
Cluster	Low coupling and high cohesion in feature space	Spatial continuity, unusual density, boundary
Outlier	Different from population or neighbors in feature space	Significant attribute discontinuity in geographic space
Association	Subset prevalence, $Pr[B \in T \mid A \in T, T : a\,transaction]$ Correlation	Spatial pattern prevalence $Pr[B \in N(A) \mid N : neighborhood]$ Cross K-Function

Table 3.4: Interest Measures of Patterns for Classical Data Mining and Spatial Data Mining

tions of nests and 3.10(b) shows the pixels with actual nests. Note the loss of information during the discretization of continuous space into pixels. Many nest locations barely fall within the pixels labeled 'A' and are quite close to other blank pixels, which represent 'no-nest'. Now consider two predictions shown in Figure 3.10(c) and 3.10(d). Domain scientists prefer prediction 3.10(d) over 3.10(c), since the predicted nest locations are closer on average to some actual nest locations. However, the classification accuracy measure cannot distinguish between 3.10(c) and 3.10(d) since spatial accuracy is not incorporated in the classification accuracy measure. Hence, there is a need to investigate proper measures for location prediction to improve spatial accuracy.

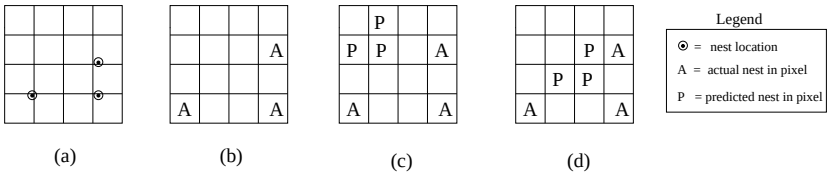


Figure 3.10: (a)The actual locations of nests, (b)Pixels with actual nests, (c)Location predicted by a model, (d)Location predicted by another model. Prediction(d) is spatially more accurate than (c).

• *Effective visualization of spatial relationships*
Visualization in spatial data mining is useful to identify interesting spatial patterns. As we discussed in Section 3.2, the data inputs of spatial data mining have both spatial and non-spatial features. To facilitate the visualization of spatial relationships, research is needed on ways to represent both spatial and non-spatial features.

For example, many visual representations have been proposed for spatial outliers. However, we do not yet have a way to highlight spatial outliers within visualizations of spatial relationships. For instance, in variogram cloud (Figure 3.5 (a)) and scatterplot (Figure 3.6 (b)) visualizations, the spatial relationship between a single spatial outlier and its neighbors is not obvious. It is necessary to transfer the information back to the original map in geographic

space to check neighbor relationships. As a single spatial outlier tends to flag not only the spatial location of local instability but also its neighboring locations, it is important to group flagged locations and identify real spatial outliers from the group in the post-processing step.

• *Improving computational efficiency*
Mining spatial patterns is often computationally expensive. For example, the estimation of the parameters for the spatial autoregressive model is an order of magnitude more expensive than that for the linear regression in classical data mining. Similarly, co-location mining algorithm is more expensive than the apriori algorithm for classical association rule mining [Agrawal & Srikant1994]. Research is needed to reduce the computational costs of spatial data mining algorithms by a variety of approaches including the classical data mining algorithms as potential filters or components.

• *Preprocessing spatial data*
Spatial data mining techniques have been widely applied to the data in many application domains. However, research on the preprocessing of spatial data has lagged behind. Hence, there is a need for preprocessing techniques for spatial data to deal with problems such as treatment of missing location information and imprecise location specifications, cleaning of spatial data, feature selection, and data transformation.

3.7 Summary

In this chapter we have presented the features of spatial data mining that distinguish it from classical data mining in the following four categories: input, statistical foundation, output, and computational process as shown in Table 3.5. We have discussed major research accomplishments and techniques in spatial data mining, especially those related to four important output patterns: predictive models, spatial outliers, spatial co-location rules, and spatial clusters. We have also identified research needs for spatial data mining.

3.8 Acknowledgments

This work was supported in part by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014, the content of which does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

We are particularly grateful to our collaborators Prof. Vipin Kumar, Prof. Paul Schrater, Dr. Sanjay Chawla, Dr. Chang-Tien Lu, Dr. Weili Wu, and Prof. Uygar Ozesmi for their various contributions. We also thank Xiaobin Ma, Hui Xiong, Jin Soung Yoo, Qingsong Lu, Baris Kazar, and anonymous reviewers

	Classical Data Mining	Spatial Data Mining
Input	Simple types Explicit relationship	Complex types Implicit relationships
Statistical Foundation	Independence of samples	Spatial autocorrelation
Output	Set-based interest measures e.g., classification accuracy	Spatial interest measures, e.g., spatial accuracy
Computational Process	Combinatorial optimization, Numerical Algorithms	Computational efficiency opportunity Spatial autocorrelation, plane-sweeping New complexity: SAR, co-location

Table 3.5: Difference between Classical Data Mining and Spatial Data Mining

for their valuable feedbacks on early versions of this chapter. We would also like to express our thanks to Kim Koffolt for improving the readability of this chapter.

Bibliography

[Agrawal & Srikant1994] Agrawal, R., and Srikant, R. 1994. Fast Algorithms for Mining Association Rules. In *Proc. of Very Large Databases*.

[Anselin1988] Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht, Netherlands: Kluwer.

[Anselin1994] Anselin, L. 1994. Exploratory Spatial Data Analysis and Geographic Information Systems. In Painho, M., ed., *New Tools for Spatial Analysis*, 45–54.

[Anselin1995] Anselin, L. 1995. Local Indicators of Spatial Association: LISA. *Geographical Analysis* 27(2):93–115.

[Barnett & Lewis1994] Barnett, V., and Lewis, T. 1994. *Outliers in Statistical Data*. John Wiley, 3rd edition edition.

[Besag1974] Besag, J. 1974. Spatial Interaction and Statistical Analysis of Lattice Systems. *Journal of Royal Statistical Society: Series B* 36:192–236.

[Bolstad2002] Bolstad, P. 2002. *GIS Foundamentals: A Fisrt Text on GIS*. Eider Press.

[Cressie1993] Cressie, N. 1993. *Statistics for Spatial Data (Revised Edition)*. New York: Wiley.

[Han, Kamber, & Tung2001] Han, J.; Kamber, M.; and Tung, A. 2001. Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, J., eds., *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis.

[Hawkins1980] Hawkins, D. 1980. *Identification of Outliers*. Chapman and Hall.

[Jain & Dubes1988] Jain, A., and Dubes, R. 1988. *Algorithms for Clustering Data*. Prentice Hall.

[Jhung & Swain1996] Jhung, Y., and Swain, P. H. 1996. Bayesian Contextual Classification Based on Modified M-Estimates and Markov Random Fields. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 34(1):67–75.

[Koperski & Han1995] Koperski, K., and Han, J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In *Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66.*

[Li1995] Li, S. 1995. A Markov Random Field Modeling. *Computer Vision.*

[Morimoto2001] Morimoto, Y. 2001. Mining Frequent Neighboring Class Sets in Spatial Databases. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

[Quinlan1993] Quinlan, J. 1993. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

[Roddick & Spiliopoulou1999] Roddick, J.-F., and Spiliopoulou, M. 1999. A Bibliography of Temporal, Spatial and Spatio-Temporal Data Mining Research. *SIGKDD Explorations 1(1): 34-38 (1999).*

[Shekhar & Chawla2003] Shekhar, S., and Chawla, S. 2003. Spatial Databases: A Tour. *Prentice Hall (ISBN 0-7484-0064-6).*

[Shekhar & Huang2001] Shekhar, S., and Huang, Y. 2001. Co-location Rules Mining: A Summary of Results. In *Proc. of the 7th Int'l Symp. on Spatial and Temporal Databases.*

[Shekhar *et al.*2002] Shekhar, S.; Schrater, P. R.; Vatsavai, R. R.; Wu, W.; and Chawla, S. 2002. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transaction on Multimedia* 4(2).

[Shekhar, Lu, & Zhang2003] Shekhar, S.; Lu, C.; and Zhang, P. 2003. A Unified Approach to Detecting Spatial Outliers. *GeoInformatica* 7(2).

[Solberg, Taxt, & Jain1996] Solberg, A. H.; Taxt, T.; and Jain, A. K. 1996. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transaction on Geoscience and Remote Sensing* 34(1):100–113.

[Tobler1979] Tobler, W. 1979. *Cellular Geography, Philosophy in Geography.* Dordrecht, Reidel: Gale and Olsson, Eds.

[Warrender & Augusteijn1999] Warrender, C. E., and Augusteijn, M. F. 1999. Fusion of image classifications using Bayesian techniques with Markov rand fields. *International Journal of Remote Sensing* 20(10):1987–2002.

[Zhang *et al.*2003] Zhang, P.; Huang, Y.; Shekhar, S.; and Kumar, V. 2003. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries. In *Proc. of the 8th Intl. Symp. on Spatial and Temporal Databases.*